⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平2－158871

⑤Int.Cl.⁵	識別記号	庁内整理番号	㊸公開　平成2年(1990)6月19日
G 06 F 15/40	500 T	7313－5B	

審査請求　未請求　請求項の数 2　（全4頁）

㊹発明の名称　文書分類装置

㉑特　願　昭63－312107
㉒出　願　昭63(1988)12月12日

㉗発明者　森田　哲也　東京都大田区中馬込1丁目3番6号　株式会社リコー内
㉑出願人　株式会社リコー　東京都大田区中馬込1丁目3番6号
㉔代理人　弁理士　香取　孝雄　外1名

明　細　書

1．発明の名称

文書分類装置

2．特許請求の範囲

1．文書データベースにおけるキーワードの出現頻度値を用いて計算される各キーワードの自己情報量を保持するキーワード情報量記憶手段と、

前記キーワードの自己情報量を用いて各文書ごとの概念特徴量を求める概念特徴抽出手段と、

文書間の該概念特徴量の差に応じて文書間の距離を求める文書間距離計算手段とを有し、

該文書間距離計算手段は、前記文書間の距離によって文書の分類を行なうことを特徴とする文書分類装置。

2．文書データベースにおいて使用されるシソーラスのキーワード分類項目ごとにキーワードの出現頻度値を用いて計算されるキーワードの自己情報量を保持するキーワード情報量記憶手段と、

各キーワード分類項目ごとの該キーワード情報量の総和をベクトル化したものを概念特徴量として求める概念特徴抽出手段と、

文書間の該概念特徴量の差に応じて文書間の距離を求める文書間距離計算手段とを有し、

該文書間距離計算は、前記文書間の距離によって文書の分類を行なうことを特徴とする文書分類装置。

3．発明の詳細な説明

［産業上の利用分野］

本発明は文書分類装置、とくに、文書に含まれるキーワードに基き文書の概念特徴量を求め、概念特徴量により文書を分類する文書分類装置に関する。

［従来の技術］

文書をあらかじめ設定した分野へ自動的に分類するためカイ自乗値を用いてキーワードの偏りを調べ、文書を分類する方式が知られている。このような分類方式を記載したものとして、田村他「統計的手法による文書自動分類」（情報処理35回全国大会論文集、1987年）、および林知己夫「数量化の方法」（東洋経済新聞社、1974年）が

ある。

カイ自乗検定はキーワードの出現頻度の分野による偏りを示す指標としてカイ自乗値を求め文書を分類するものである。カイ自乗値は、各キーワードの出現頻度値と各分野ごとの総キーワード数が独立事象であると仮定した場合のキーワードの出現頻度値を理論度数とし、実測値との差を求め正規化したものである。

上記の文献①はカイ自乗検定を用いて文書をあらかじめ設定した分野へ自動的に分類する方式について述べたものである。この方式は、キーワードの出現頻度の偏りを用いるために、あらかじめ大量の標本データを分野別に分類してカイ自乗値を計算し、分類用データを用意しておく必要がある。

文献②もやはりカイ自乗値を用いる統計的手法の一つであり、複数の分野間の相関を見るための方式である。

［発明が解決しようとする課題］

上記の文献①②に記載された方式は、標本データの分類にはやはり人手による作業が必要となる。したがって、人手による分類のばらつきや不適切さが介入するという問題がある。

また、後者は分類用の軸を決定するのが難しいという問題がある。

本発明は上記の問題点を解決するために、文書に含まれるキーワードの頻度値から各文書の概念特徴量を求め、これに応じて文書を分類する文書分類装置を提供することを目的とする。

［課題を解決するための手段］

上記目的を達成するために、本発明によれば、文書データベースにおけるキーワードの出現頻度値を用いて計算される各キーワードの自己情報量を保持するキーワード情報量記憶手段と、キーワードの自己情報量を用いて各文書ごとの概念特徴量を求める概念特徴抽出手段と、文書間の概念特徴量の差に応じて文書間の距離を求める文書間距離計算手段とを有する。

文書間距離計算手段は、文書間の距離によって文書の分類を行う。

［作　用］

本発明によれば、キーワード情報量記憶手段が文書データベース等のキーワード出現頻度により、所定の計算を行って各キーワードの自己情報量を求め、概念特徴抽出手段が自己情報量より所定の計算により各文書の概念特徴量を求め、文書間距離計算手段が概念特徴量の差に応じて文書の分類を行なう。以上のようにキーワードの頻度より各手段の計算処理を通して、自動的に文書が分類されるので、従来の人手作業が不要となり、ばらつきのない、概念量による文書分類が構築できる。

［実施例］

本発明の実施例を図面を用いて具体的に説明する。

本発明による文書分類装置の一実施例が図に示されている。

キーワード情報量記憶部1は入力される未登録文書Qよりキーワードを抽出し、後述のようにその出現頻度よりキーワードの出現確率を求め、その対数値をキーワード情報量Iとして記憶する。概念特徴抽出部2はキーワード情報量記憶部1よりキーワード情報量Iを入力し、その総和を文書Qの概念特徴量C(q)として出力する。文書間距離計算部3は概念特徴抽出部2より各文書の概念特徴量C(q)を入力して記憶し、2つの文書間の概念距離を求めて、概念距離の近い文書をクラスタ（分類）して、各種の分類を文書データベース4に格納する。各機能部は、各部の生成したデータを転送するデータバスa～cによって接続されている。

一般にシソーラス等のキーワード集に登録されているキーワードは、それらが現われる文書数や全文書における延べ出現回数等によって各キーワードの出現頻度を定義できる。いまキーワードKEYiの出現頻度を全キーワード数で正規化したP_iをキーワードKEYiの出現確率とすると、キーワードに出現確率P_iを対応させるシステムは完全事象系となり以下のように表せる。

$$\begin{bmatrix} KEY_1 & KEY_2 & \cdots & KEY_i & \cdots & KEY_n \\ P_1 & P_2 & & P_i & & P_n \end{bmatrix}$$

ただし、$\sum_{i=1}^{n} P_i = 1$ である。

ここで、KEYiの自己情報量 I (KEYi) は次式で表せる。

$$I(KEYi) = -\log Pi \quad \cdots (1)$$

また自己情報量は加法性を保つため、KEYiとKEYjの持つ合成情報量は、次式で表わされる。

$$I(KEYi, KEYj) = I(KEYi) + I(KEYj)$$
$$= -\log Pi - \log Pj \quad \cdots (2)$$

キーワード情報量記憶部1は、文書データベース4への未登録文書Qを概念特徴抽出部2を介してデータパスaより入力し、文書Qの各キーワードを抽出し、その出現確率KEYiを求め、(1)式によりキーワードの自己情報量 I (KEYi) を計算して保持する。シソーラスが用意されているときは、シソーラスのキーワード分類項目ごとにキーワードの出現確率を求め、(1)式により自己情報量を計算できる。

ある文 Qのキーワード集合をqとしその概念特徴量をC(q)と表すと、

$$C(q) = -\sum_{KEYi \in q} \log Pi \quad \cdots (3)$$

で与えられる。

また既存の分類項目を持つシソーラスにおいては概念特徴量をベクトルとして扱うことができる。最も単純な例として、M個の分類項目を持つシソーラスではM次元のベクトルCVを考える。今、R番目の分類項目に属するキーワードの集合をrとすると、文書Qの概念特徴量ベクトルCV(q) のR要素CVr(q)は、

$$CVr(q) = -\sum_{i \in q \cap i \in r} \log Pi \quad \cdots (4)$$

ただし、$i \in q \cap i \in r$ はキーワードiが文書Q中に含まれ、かつR番目の分類項目中に含まれている場合のPiの総和を計算することを意味する。

キーワード情報量記憶部1から文書Qの各キーワードの自己情報量Iを入力し、概念特徴抽出部2は、(3) 式または(4) 式を用いて、概念特徴量C(q)またはCVr(q)を計算し、データパスbより文書間距離計算部3に出力する。

(3) 式によって求められた概念情報量はある文書のもつキーワード情報量の和であり、その文書に付加された自己情報量の大きさを示しているだけである。この場合の概念情報量は、文書データベースの検索時における当該文書の分離度の高さ（同定しやすさ）を表す。このような分離度の高さによって文書を分類することも可能である。

しかし、通常は文書の内容によって既存の分類項目等に分類する用途が考えられる。そのような場合、(4)式の概念特徴量ベクトルを用いる。一般にM個の分類項目によってデータベースはM次元の概念空間を構成すると考えられる。従ってこのようなデータベース中の文書の持つ概念は、M個の特徴パラメータからなるM次元ベクトルとして表現できる。また任意の2つの概念特徴量ベクトルの距離が計算できるため、ある文書のある分類への帰属度や2つの文書間の概念的距離等が求められる。

例えば、CV(q) という概念特徴量ベクトルを持つ文書が、キーワード集合kをもつ分類Kに帰属する度合をINC(k,q)とすると、

$$INC(k,q) = CVk(q) \Big/ \sum_{r=1}^{M} CVr(q) \quad \cdots (5)$$

で与えられる。

また、CV(s) 、CV(t) という概念特徴量ベクトルを持つ2つの文書間の概念距離をD(s,t)とし例えば市街地距離で計算すると、

$$D(s,t) = \sum_{r=1}^{M} |CVr(s) - CVr(t)| \quad \cdots (6)$$

で与えられる。

文書間距離計算部3は概念特徴量C(q)またはCV(q) を入力し、(5) 式で示した計算を行なうことにより、未分類の文書の属すべき分類を決定でき、また (6)式を用いると、概念距離の近い文書群によっていくつかの分類を構成できる。文書間距離計算部3は文書Qの分類を文書データベース4に入力する。このとき生成される分類は、既存

のいくつかの分類項目の概念を結合した合成概念になるため、既存の分類項目に縛われない文書概念自体に指向した新しい分類体系を自然に構築していく。

(6)式を用いた同類文書の分類方法について具体的に説明する。

前述のように既存の分類項目に対して文書分類を行なう場合には、(5)式を用いて各分類Kに帰属する度合いINC(k,q)を求めればよい。さらに概念特徴量ベクトルを用いると、既存の分類項目を用いて新しい分類体系を構築することが可能となる。

まず、分類しようとする全ての文書について各文書間の概念距離Dを求める。次に全ての文書の中から任意に1文書(文書Sとする)を選択し、その文書との概念距離が所定のしきい値より小さい、すなわちその文書と概念的に近い文書を抽出する。抽出された文書Tの集合を式で表現すると、文書S、Tに含まれるキーワード集合をそれぞれt、sとすれば、

(T | D(s, t) < θ)

(ただし、D(s, s) = 0は(6)より明らかであり、文書Sは必ず集合Tに含まれる。)

この作業を全ての文書に対して行なうと文書数に等しい同類文書の集合が出来上がる。これら同類文書集合をその集合の要素数(文書数)に従って降順に並べ、文書数の多い順に必要な分類数だけの同類文書集合を選択する。この選択は分類数で制限しても良いし、文書数で制限しても良い。分類可能な数の最大値は文書数である。この場合各分類に含まれる文書数は1であるが、このような分類が最適となる場合もあってしかるべきである。

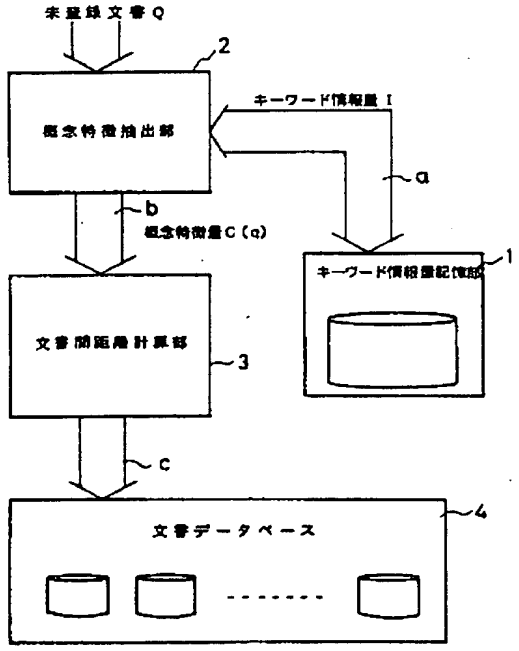
[発明の効果]

本発明によれば、キーワード抽出、または既存のキーワード集の分類を用いて概念特徴量を計算できるため、未登録文書の分類の前に評価用データを作成する必要がない。

概念距離の近い文書群によって分類を構成するため、既存の分類項目に縛われない文書概念自体に指向した新しい分類体系を自然に構築していくという優れた効果がある。

4. 図面の簡単な説明

図は本発明の文書分類装置の一実施例を示す機能ブロック図である。



主要部分の符号の説明

1…キーワード情報量記憶部、
2…概念特徴抽出部、
3…文書間距離計算部、
4…文書データベース。

特許出願人　株式会社リコー

代理人　香取　孝雄
丸山　隆夫